I3D

2019/05/25 김민지

Abstract

- How Current Architectures fare on the task of **action classification** on Kinetics dataset?
- How much **performance improves on the smaller benchmark datasets** after pre-training on Kinetics.

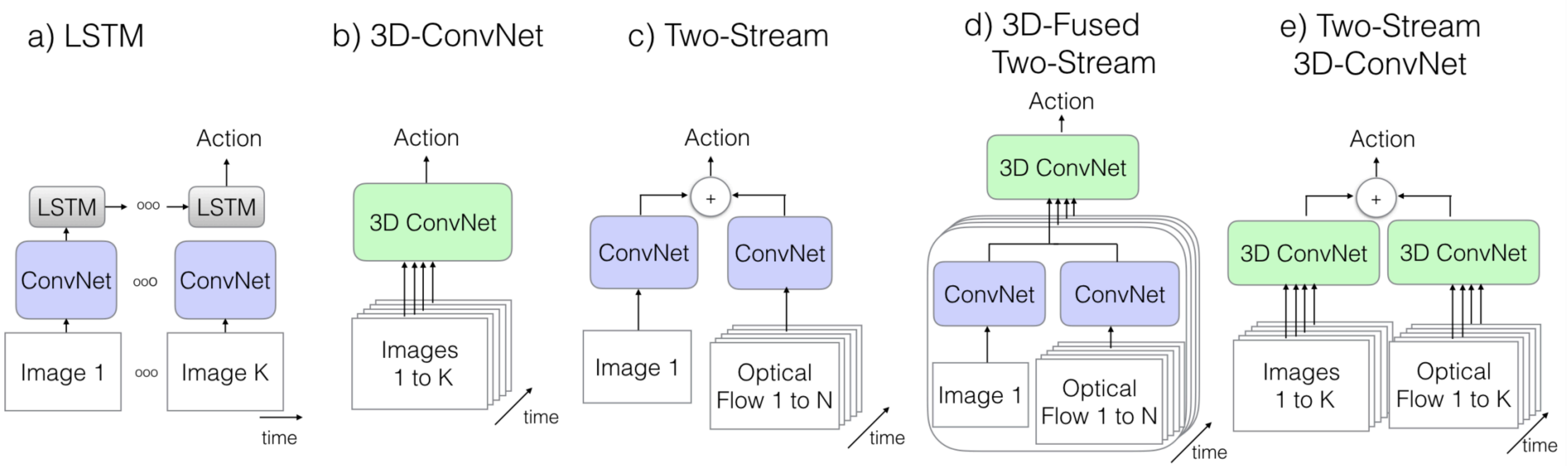
Abstract

- Two-Stream Inflated 3D ConvNet(I3D)
  : To learn **seamless** spatio-temporal feature extractors
  + leveraging successful ImageNet architecture **designs & parameters**.

Introduction

- Most popular benchmarks for action recognition are small!
  (the order of 10k videos)
  Kinetics → HMDB-51, UCF-101. (two orders larger)

Action Classification Architectures



- The Old 1: ConvNet + LSTM
- The Old 2: 3D ConvNet
- The Old 3: Two-Stream Networks

- The new: Two-Stream Inflated 3D ConvNets

[Joao Carreira](#) et al., **Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,** CVPR, 2018

Two-Stream Inflated 3D ConvNets

1. Inflating 2D ConvNets into 3D

- Convert 2D Classification **models** into 3D ConvNets.

- Starting with **2D architecture**

  → Inflating all the filters and pooling kernels(temporal dimension)

- N x N filters becomes N x N x N

2. Bootstrapping 3D filters from 2D filters.

- Bootstrap **parameters** from the pre-trained ImageNet models.

Boring-video fixed point

- The pooled activations on a boring video
  == on the original single-image input. (by linearity)

➔ the overall network response respects the boring-video fixed point.

3. Pacing receptive field growth in space, time and network depth

- How to inflate pooling along the time
- How to set conv/pooling temporal stride?

- Virtually all image models treat the two spatial dimensions equally

- In Time dimension?
- Grows too quickly in time → conflate edges from different objects, breaking early feature detection
- Grows too slowly → may not capture scene dynamics well

3. Pacing receptive field growth in space, time and network depth

- 2D inception v-1
- 첫번째 conv: stride 2,
  4개의 Max-pool: stride 2,
  7x7 average-pool

- Not perform temporal pooling in the first two max-pool(1 x 3 x 3), 다른 max-pool에서는 symmetric하게 kernel, strides 이용
- 25fps, 64-frame snippets, testing with whole videos, averaging predictions temporally.

3. Pacing receptive field growth in space, time and network depth

Two 3D Streams(feat. Optical Flow)

- 3D Conv도 RGB에서 motion feature 바로 얻을 수 있지만, 여전히 feedforward computation
- Optical flow algorithm을 이용하면 어느정도 recurrent

Architecture	UCF-101			HMDB-51			Kinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	63.3	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	56.1	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	62.2	52.4	65.6
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	–	–	67.2
(e) Two-Stream I3D	**84.5**	**90.6**	**93.4**	**49.8**	**61.9**	**66.4**	**71.1**	**63.4**	**74.2**

- Trained the two networks separately and averaged their predictions at test time.

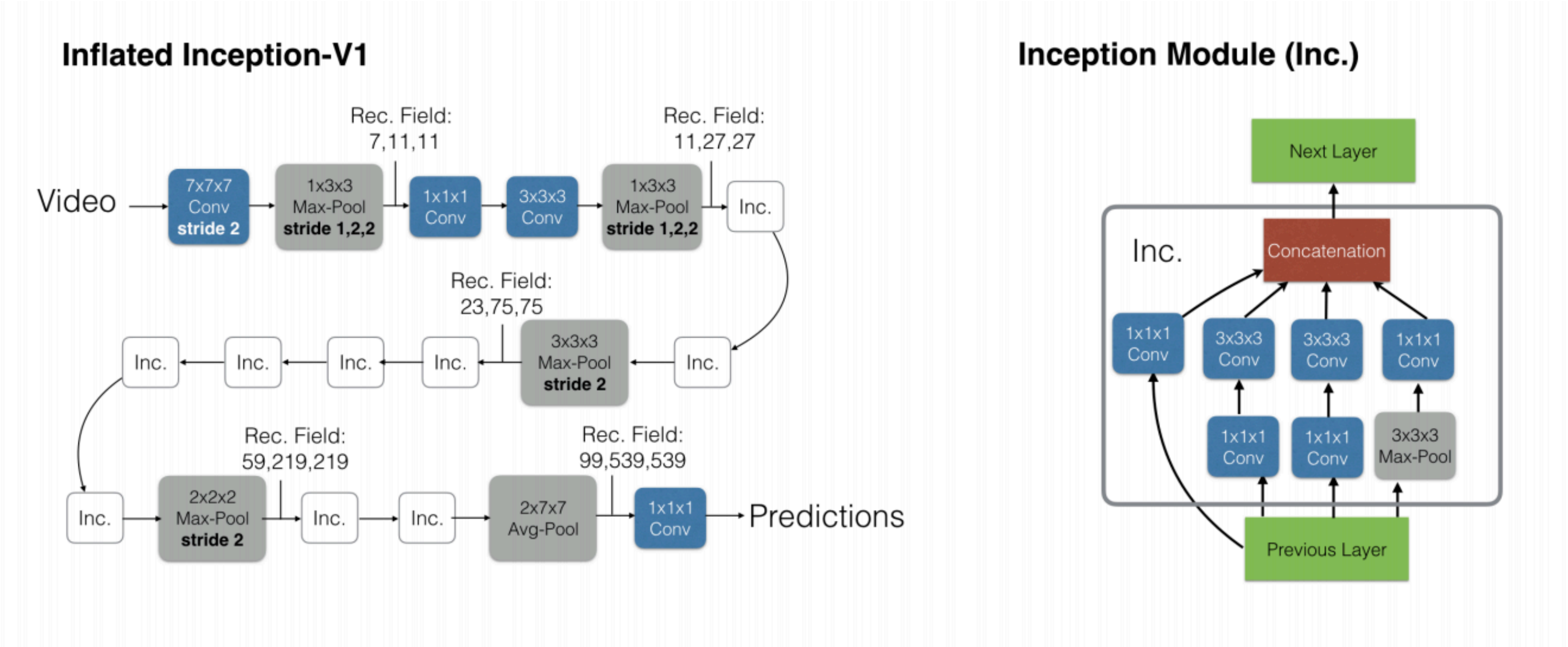


Figure 3. The Inflated Inception-V1 architecture (left) and its detailed inception submodule (right). The strides of convolution and pooling operators are 1 where not specified, and batch normalization layers, ReLu's and the softmax at the end are not shown. The theoretical sizes of receptive field sizes for a few layers in the network are provided in the format "time,x,y" – the units are frames and pixels. The predictions are obtained convolutionally in time and averaged.

Joao Carreira et al., **Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,** CVPR, 2018

The Kinetics Human Action Video Dataset

Kinetics dataset

- Focused on human actions (not activities or events)
- Person Actions (drawing, drinking …)
- Person-Person Actions (hugging, kissing, shaking hands, …)
- Person-Object Actions(opening present, washing dishes, ..)

→Swimming vs Washing dishes
→Temporal reasoning vs Emphasis on object

- 400 classes x 400 clips each class x 10s per clips

Experimental Comparison of Architectures

Architecture	UCF-101			HMDB-51			Kinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	63.3	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	56.1	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	62.2	52.4	65.6
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	–	–	67.2
(e) Two-Stream I3D	**84.5**	**90.6**	**93.4**	**49.8**	**61.9**	**66.4**	**71.1**	**63.4**	**74.2**

Table 2. Architecture comparison: (left) training and testing on split 1 of UCF-101; (middle) training and testing on split 1 of HMDB-51; (right) training and testing on Kinetics. All models are based on ImageNet pre-trained Inception-v1, except 3D-ConvNet, a C3D-like [31] model which has a custom architecture and was trained here from scratch. Note that the Two-Stream architecture numbers on individual RGB and Flow streams can be interpreted as a simple baseline which applies a ConvNet independently on 25 uniformly sampled frames then averages the predictions.

Architecture	Kinetics			ImageNet then Kinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	53.9	–	–	63.3	–	–
(b) 3D-ConvNet	56.1	–	–	–	–	–
(c) Two-Stream	57.9	49.6	62.8	62.2	52.4	65.6
(d) 3D-Fused	–	–	62.7	–	–	67.2
(e) Two-Stream I3D	**68.4** (88.0)	**61.5** (83.4)	**71.6** (90.0)	**71.1** (89.3)	**63.4** (84.9)	**74.2** (91.3)

Table 3. Performance training and testing on Kinetics with and without ImageNet pretraining. Numbers in brackets () are the Top-5 accuracy, all others are Top-1.

Joao Carreira et al., **Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,** CVPR, 2018

Architecture	UCF-101			HMDB-51		
	Original	Fixed	Full-FT	Original	Fixed	Full-FT
(a) LSTM	81.0 / 54.2	88.1 / 82.6	91.0 / 86.8	36.0 / 18.3	50.8 / 47.1	53.4 / 49.7
(b) 3D-ConvNet	– / 51.6	– / 76.0	– / 79.9	– / 24.3	– / 47.0	– / 49.4
(c) Two-Stream	91.2 / 83.6	93.9 / 93.3	94.2 / 93.8	58.3 / 47.1	66.6 / 65.9	66.6 / 64.3
(d) 3D-Fused	89.3 / 69.5	94.3 / 89.8	94.2 / 91.5	56.8 / 37.3	69.9 / 64.6	71.0 / 66.5
(e) Two-Stream I3D	93.4 / 88.8	97.7 / 97.4	98.0 / 97.6	66.4 / 62.2	79.7 / 78.6	81.2 / 81.3

Table 4. Performance on the UCF-101 and HMDB-51 test sets (split 1 of both) for architectures starting with / without ImageNet pretrained weights. Original: train on UCF-101 or HMDB-51; Fixed: features from Kinetics, with the last layer trained on UCF-101 or HMDB-51; Full-FT: Kinetics pre-training with end-to-end fine-tuning on UCF-101 or HMDB-51.

Model	UCF-101	HMDB-51
Two-Stream [27]	88.0	59.4
IDT [33]	86.4	61.7
Dynamic Image Networks + IDT [2]	89.1	65.2
TDD + IDT [34]	91.5	65.9
Two-Stream Fusion + IDT [8]	93.5	69.2
Temporal Segment Networks [35]	94.2	69.4
ST-ResNet + IDT [7]	94.6	70.3
Deep Networks [15], Sports 1M pre-training	65.2	-
C3D one network [31], Sports 1M pre-training	82.3	-
C3D ensemble [31], Sports 1M pre-training	85.2	-
C3D ensemble + IDT [31], Sports 1M pre-training	90.1	-
RGB-I3D, Imagenet+Kinetics pre-training	95.6	74.8
Flow-I3D, Imagenet+Kinetics pre-training	96.7	77.1
Two-Stream I3D, Imagenet+Kinetics pre-training	**98.0**	80.7
RGB-I3D, Kinetics pre-training	95.1	74.3
Flow-I3D, Kinetics pre-training	96.5	77.3
Two-Stream I3D, Kinetics pre-training	97.8	**80.9**

Table 5. Comparison with state-of-the-art on the UCF-101 and HMDB-51 datasets, averaged over three splits. First set of rows contains results of models trained without labeled external data.

Joao Carreira et al., **Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,** CVPR, 2018